The Hitchhiker's Guide to Mechanistic Interpretability of Vision-Language-Action Models

Aryan Goyal

Independent Researcher · Carnegie Mellon University · aryangoyal7.github.io

A map of where the LLM interpretability toolkit applies to VLAs unchanged, where it needs adaptation, and where it must be rebuilt.

7 toolkit assumptions across **7 VLA architecture classes**.

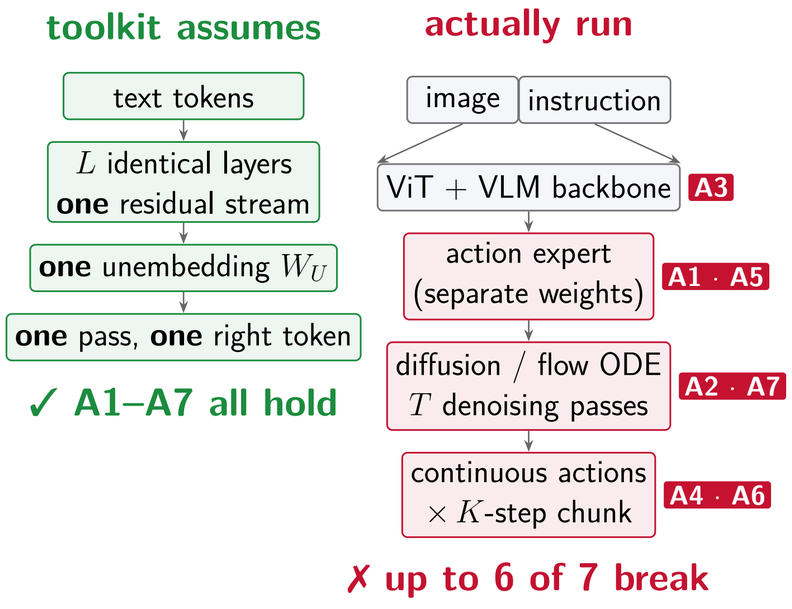
❶ The problem

The LLM interpretability toolkit rests on architectural assumptions: a **single residual stream**, a **shared unembedding matrix**, **one forward pass** per decision. VLAs violate between **one and six** of these, in different ways depending on the architecture. Each violation needs either a **non-trivial fix** or a **new tool**.

This positioning paper establishes that map for **all VLA architecture classes**: which assumptions hold, which break, and what to use instead.

Example: Direct Logit Attribution on a cross-attention VLA runs without error and **silently drops the entire visual pathway**.

What the toolkit assumes vs. what a VLA is



❷ The seven assumptions (A1–A7)

The tool assumes	A VLA breaks it because
A1 one residual stream	cross-attention & action experts write *off-stream*
A2 shared unembedding W_U	continuous action heads have *no* W_U at all
A3 every layer is alike	vision + backbone + action head = 3 different networks
A4 one correct token	robots have many valid trajectories
A5 one uniform architecture	experts route token types to different weights
A6 low superposition	256-bin action codes densely superpose
A7 one pass = one decision	diffusion / flow / chunking take many passes

❸ Find your architecture, read off what breaks

VLA type (example)	A1	A2	A3	A4	A5	A6	A7	breaks
I Discrete-token (RT-2, OpenVLA)	✓	✓	∼	✓	✓	✗	✓	1
II Continuous-regression (GR-1)	✓	✗	✓	✓	✓	✗	✓	2
III Diffusion (Octo)	✗	✗	✓	✓	∼	✗	✗	4
IV Flow-matching (π_0)	✗	✗	✓	∼	✗	✗	✗	5
V Cross-attn fusion (RoboFlamingo)	✗	✗	∼	✓	✗	✗	✓	4
VI Heterogeneous-expert (SpatialVLA)	✓	✓	✓	✓	✗	✗	✓	2
VII Temporal-chunking (ACT)	✗	✗	✓	✗	✗	✗	✗	6

✓ holds ∼ partial ✗ breaks

The toolkit fully applies **only** to Type I. **Flow-matching breaks 5 of 7**, and flow-matching models (π_0, $\pi_{0.5}$) now lead generalist robot control.

❹ Five silent failures and their fixes

Silent failure	Fix
F1 cross-attention carries fraction α of the signal, so every DLA score is deflated by $(1-\alpha)$	trust *rankings*, never absolute numbers
F2 with no W_U, text-side DLA scores a readout the model never uses	supervised probe readout (**ProbeDLA**)
F3 SVD returns variance axes, not action axes ($100\times$ scale gap drowns gripper dims)	rank directions by how well they *predict* the action
F4 token-flip metrics miss behaviour: bin $31\to32$ "flips" yet acts identically	score in *action space*: KL, rollout success
F5 sequence-mean ablation is out-of-distribution across visual / language / action positions	*position-preserving* mean ablation

Why rollouts cannot validate your analysis

Behavioural cloning compounds per-step error: a policy that is wrong by ε per step can accumulate regret up to εH^2 over an H-step rollout (Ross & Bagnell 2010). A VLA executes $H{=}100$–200 physical steps with no recovery data, and with continuous actions the compounding can become exponential (Simchowitz et al. 2025).

An intervention (an ablation, a steering vector) is judged through these same dynamics, so the rollout outcome is not confirmatory:

- a **successful** rollout does not confirm it: other components compensate;
- a **failed** rollout does not refute it: the dynamics amplify small intervention errors.

A rollout measures **behaviour**, not the correctness of an attribution.

How to use this field guide

1. **Find your VLA's row** in the map ❸.
2. **Read off which assumptions** A1–A7 it breaks.
3. **Apply the fixes** ❹: ProbeDLA, action-space metrics, position-preserving ablation, per-step attribution.
4. **Validate in action space**, not by rollout alone.

The takeaway

Every architecture beyond Type I needs at least one non-trivial fix, and the field is moving toward flow and diffusion action heads, where the most tools break.

Before running an LLM tool on a VLA, check which of its assumptions still hold.

Open for collaborations:
aryangoyal7.github.io

Paper
openreview.net

ICML 2026 Workshop on Mechanistic Interpretability

Website / contact
aryangoyal7.github.io